On the Definition of Objective Probabilities by Empirical Similarity*

Itzhak Gilboa†, Offer Lieberman‡, and David Schmeidler§

May 2007

Abstract

We suggest to define objective probabilities by similarity-weighted empirical frequencies, where more similar cases get a higher weight in the computation of frequencies. This formula is justified intuitively and axiomatically, but raises the question, which similarity function should be used? We propose to estimate the similarity function from the data, and thus obtain objective probabilities. We compare this definition to others, and attempt to delineate the scope of situations in which objective probabilities can be used.

1 Definitions of Probability

How should we assign probabilities to events? What is the meaning of a statement of the form, "Event A will occur with probability p"? Or, to be more cautious, under which conditions can we assign probabilities to events, and under which conditions do possible meanings of the term apply?

*We wish to thank Gabi Gayer, Jacob Leshno, Arik Roginsky, and Idan Shimony for comments and references. This project was supported by the Pinhas Sapir Center for Development and Israel Science Foundation Grants Nos. 975/03 and 355/06.

†Tel-Aviv University, HEC, and Cowles Foundation, Yale University. igilboa@tau.ac.il

‡University of Haifa. offerl@econ.haifa.ac.il

§Tel-Aviv University and The Ohio State University. schmeid@tau.ac.il

To address these questions we consider a few examples. For concreteness, we embed these examples in decision problems, and one may further suppose that the decision makers in question are interested in assigning probabilities to events in order to subsequently maximize expected utility. Yet, our focus is on the concept of "probability" as such. In discussing these problems, observe also that our approach is epistemological rather than ontological. We do not purport to discuss the "true" nature of probability, but only the notion of probability insomuch as it can be measured and quantified. We are interested in the type of circumstances in which a statement "Event A will occur with probability p" can be made, and the meaning that should then be attached to such a statement.

Example 1 A coin is about to be tossed. Sarah is offered a bet on the outcome of the toss. She wonders what is the probability of the event "The coin lands on Head".

Example 2 John normally parks his car on the street. He is offered an insurance policy that will cover theft. To decide whether he should buy the policy, he wonders what is the probability that his car will be stolen during the coming year.

Example 3 Mary has to decide whether to undergo a medical operation that is supposed to improve her quality of life, but that may also involve serious risks. Trying to make a rational decision, Mary asks her physician what is the probability of various events, such as death.

Example 4 George considers an investment opportunity, and he figures out that the investment will not be very successful if there is another war in the Middle East over the next year. He then attempts to assess the probability of such a war in this time frame.

The term "probability" has various meanings and definitions. Among these, at least three seem to be widely accepted:

The "Classical" approach suggests that all possible outcomes have the same probability. This approach has been used in early writings on games of

chance, and it has been explicitly formulated by Laplace as a principle, later dubbed the "Principle of Insufficient Reason" or the "Principle of Indifference".

The "Frequentist" approach offers the empirical frequency of an event in past observations as a definition of its probability. Bernoulli's (1713) law of large numbers guarantees that independent and identical repetitions of an experiment will result, with probability 1, in a relative frequency of occurrence of an event that converges to the event's probability. In fact, this limit relative frequency is often used as an intuitive definition of probability. The relative frequency in a finite sample can thus be a good estimate, or even a definition of the probability of the event.

The "Subjective" approach views probability as a numerical measure of degree of belief that is constrained to satisfy certain conditions (or "axioms"). Subjective probability has been discussed from the very first days of probability theory, and it is used already by Pascal in his famous "wager". Ramsey (1931), de Finetti (1937), and Savage (1954) have promoted it, and suggested axioms on observed behavior, that would necessitate the existence and uniqueness of a subjective probability measures. Specifically, Savage (1954) provided a set of axioms on choices between alternative courses of actions, which imply that the decision maker behaves as if she wished to maximize the expectation of a certain function with respect to a certain probability measure. Interpreting the function as "utility" and the measure as "subjective probability", his theorem provides a behavioral definition of subjective probability, coupled with the principle of expected utility maximization.

We now turn to examine how each of these three approaches deals with the four examples described above.

1.1 The Classical Approach

The classical approach applies in Example 1. When Sarah is considering the bet on the outcome of the coin toss, she might, in the absence of any other information, assign a probability of 50% to each possible outcome. However, the same approach does not seem to be tenable in the other examples. If John were to say, "either my car is stolen, or it isn't, hence it has probability of 50% of being stolen", he would hardly be rational. Nor can Mary or George assign 50% to surviving the operation, or to a war in the Middle East, respectively. Clearly, there is too much information in these examples to apply the principle of insufficient reason. In fact, even in the absence of information, this principle has come under attack on various grounds. For instance, it is very sensitive to the representation of the state space, especially when the latter in infinite.

Despite these attacks, in Example 1 the principle of insufficient reason is acceptable, whereas in Examples 2-4 it is completely inappropriate. One may wonder what are precisely the features of Example 1 that distinguish it from the others in this respect. That is, one may wish to delineate the scope of applicability of the principle of insufficient reason. Here we merely conclude that this approach for the assignment of probabilities is not very useful in most real life decision problems.

1.2 The Frequentist Approach

The frequentist approach appears to be more promising. Like the classical approach, it can deal with the coin problem in Example 1: if one has many observations of tosses of the same coin in the past, conducted under similar conditions, one may take the observed relative frequency as a definition of the probability of the coin landing on Head. Indeed, it is quite possible that this approach will coincide with the principle of insufficient reason, as will be the case if the coin is fair. But the relative frequency approach would apply equally well also if the coin is not fair. This approach does not assume any

symmetries in the problem, and it is therefore robust to alternative representations of the state space. The frequentist approach deals with Example 2 in basically the same way as with Example 1: if there are many observations of cars parked overnight, and if these observations were taken under practically the same conditions, that is, in the same neighborhood, for the same type of car, and so forth, then it makes sense to take the average rate of theft as the probability of a particular car being stolen.

How would the frequentist approach be applied to Example 3? In principle, it should follow the same pattern: Mary should ask her physician how often the operation has succeeded in the past, and use the ratio of successes to trials as the "probability" of success. But Mary may find her physician uneasy about a straightforward quote of relative frequencies. After all, the physician might say, the data were collected over a variety of individuals, who differ from Mary in many relevant ways, including age, gender, weight, blood pressure, and so forth. They were operated on in different hospitals and by different surgeons. In fact, the physician might say, since no two cases are identical, you can choose which dataset to look at, and thereby affect the "probability" you obtain. Thus, the objectivity of the empirical frequency approach is compromised by the subjectivity of the choice of the sample.

The applicability of the frequentist approach to Example 4 is even more dubious. This approach would call for the listing of past cases in which war has or has not erupted, and taking the number of wars divided by the overall number of cases to be the probability of war. One difficulty that becomes obvious in this example is the precise delineation of a "case" in time. Should we take each year to be a separate case? If so, how would we deal with a war that lasted more than one year, or with a year in which more than one war has occurred? Should we perhaps lump periods together in larger chunks? Or should we define cases as starting and ending by historically meaningful events? Clearly, splitting and merging cases will affect the relative frequencies of wars, and thereby our probability assessments.

A second difficulty that George would encounter in Example 4 is that encountered by Mary in Example 3: the choice of the dataset, or the relevant "sample" is not obvious. What should count as a case, relevant for the relative frequency of the occurrence of wars? Should we go back to wars in the Middle East in biblical times? These might be relevant when certain geographical or strategic considerations are concerned, but their relevance seems limited, as well as our degree of confidence in their veracity. Should we perhaps restrict attention to modern times? But if so – how can one define "modern times" objectively? Does it make more sense to restrict one's attention to post-WWII period, or to rely on a larger dataset that predates WWII? Similarly, one may further wonder which other features of past cases should matter. Should one consider only cases in which the involved parties had similar military might, similar regime, or similar economic conditions? Clearly, as in Mary's medical example, George also faces a situation that is unique. History repeats itself, but never in precisely the same form, and the current case has enough features that distinguish it from all past cases. If George were to take all these considerations into account, he will end up with an empty dataset. If he ignores them completely, his dataset is large but very uneven in terms of relevance. Thus, the choice of the dataset becomes a subjective one, which ends up affecting the assessed probability.

There is yet another difficulty that is unique to George's problem: in Example 4, past cases cannot be assumed to be causally independent.[1] Thus, the relative frequency approach may be ignoring important mechanisms that are at work. Observe that, in Example 3, Mary could ignore possible causal dependencies. She could argue that the success of the operation on other patients does not directly affect its success in her case. This is clearly an

[1] We use the term "causation" in an intuitive sense. Of the various definitions of this term, some resort to probability as a primitive (cf. Pearl, 2000). Such definitions cannot be used in our context, as we are attempting to define the term "probability". However, we do not use causation as part of our suggested definition. It will only be used in the informal meta-discussion, attempting to characterize the scope of applicability of our definition.

assumption about the world. Moreover, it may not be true, if, for instance, Mary is going to be operated on by a surgeon who failed in his previous operation and may have even been sued for malpractice. Still, the independence assumption seems like a reasonable one for Mary, and it allows her to view relative frequencies as proxies for "probability". This is not the case in George's problem. Recent wars are intricately related in various causal relationships to the possible next war. There are political and military lessons that are being learned, there are goals that have and have not been obtained, and so forth. Hence, simply considering relative frequencies may be completely misleading.

1.3 The Subjective Approach

The subjective approach appears to be immune to all the difficulties mentioned above. According to this approach, probabilities are subjective, or "personal", and therefore they need not derive from past data or from perceived symmetries. Rather, they reflect intuition, and model it in a precise way. One may have a degree of belief in the eruption of war just as one has a degree of belief in a coin landing on Head, and the formal probability model can help sharpen this intuition and put it to use. Moreover, making decisions in accordance with certain sets of axioms implies that one makes decisions as if one were to use a probability measure.[2]

The subjective approach is conceptually very neat. Rather than coping with the essence of objectivity, with the meaning of factual knowledge, and with the possibility of processing data in an objective way, this approach steps

[2]Probability may be used in various decision rules, the most famous of which is expected utility maximization. But a decision maker may be following a well-defined subjective probability measure also when using other rules. Machina and Schmeidler (1992) defined and axiomatized "probabilistic sophistication", which may be defined as "having a subjective probability measure and making decisions based solely on the distributions that this probability induces". Rostek (2006) suggested axioms that imply that the decision maker has a subjective probability measure, and that she makes decisions so as to maximize the median utility with respect to that probability.

back, gives up any claim to objectivity, and rearranges its defense around universality: probability is only subjective, but, as such, it may apply to any source of uncertainty, irrespective of the amount of relevant data gathered.

However, this approach has been attacked on several grounds. Ellsberg's experiments (Ellsberg, 1961) have shown that people often behave as if they do not have a subjective probability measure that may summarize their beliefs. Several authors have also attacked the subjective approach on normative grounds. (See Shafer, 1986, and Gilboa, Postlewaite, and Schmeidler, 2006.) In particular, it has been argued that in the absence of information, it may not be rational to choose a single probability measure, a choice that is bound to be arbitrary. Moreover, the behavioral derivations of probabilistic beliefs have also been criticized on the normative appeal of their underlying axioms.

In this paper we do not take issue with the subjective approach. Rather, our focus is on the possibility of defining objective probabilities. We therefore consider Examples 3 or 4, and ask whether some intuitive notion of objective probabilities can be defined in these examples.

1.4 Extending Frequentism

Let us consider Example 3 again. The main difficulty that Mary was facing in applying the frequentist approach was that past cases differed in many ways, and that each case was basically unique. It is useful to observe that, in principle, the same objection may apply to the application of the frequentist approach in Examples 1 and 2 as well. In Example 2, for instance, one might argue that no two cars are identical, just as no two patients are in Example 3. Even in Example 1 we may have to admit that no two tosses of a coin are precisely identical. Various factors distinguish one toss from another, such as meteorites that might affect Earth's gravitational field, the mood of the person tossing the coin, and so forth.

More generally, every case is unique, if only because it can be defined

by its exact time and location. Thus, the differences between Examples 1, 2, and 3, as far as the frequentist approach is concerned, are differences of degree, not of kind. All cases are inherently unique, but in examples such as 1 and 2 one may make the simplifying assumption that a certain "experiment" was repeated many times. In other words, it is a judgment of similarity that allows the frequentist approach to be used.

This observation paves the way to a natural generalization of the frequentist approach: if cases are not identical, or if there aren't sufficiently many cases that may be assumed identical, one may bring forth the similarity between cases and use it in the definition of probability. This would be in line with Hume's (1748) focus on similarity as key to prediction. Specifically, the probability of an event can be defined by its weighted relative frequency in past cases, where each case is weighed by its similarity to the present case. Thus, a success in an operation of another patient in the past makes a success in Mary's case more likely, but the degree to which the past case matters depends on the similarity between Mary and the patient in the past observation. We devote Section 2 to a more formal description of this approach, as well as to further discussion of the similarity-weighted frequency formula and its axiomatic derivations.

1.5 Is It Objective?

The similarity-weighted frequency approach may thus overcome some of the difficulties encountered by the frequentist approach. But will we not give up objectivity in this process? A given dataset will result in a large range of possible "probabilities", depending on the similarity function that we choose to employ. If the similarity function is a matter of subjective judgment, so is the resulting probability. It would therefore appear that the similarity-weighted frequency approach has, at best, translated the question, "Which probability should we use?" to "Which similarity should we use?"

However, we maintain that this translation is a step forward. In fact,

we argue that the choice of the similarity function need not be arbitrary or subjective: we propose to estimate the similarity function from the data. The basic idea is to try explaining past data by a similarity-weighted frequency formula, and, in this context, to ask which similarity function best explains the data we have observed. Section 3 describes this estimation procedure in more detail.

We thus suggest similarity-weighted frequencies, employing the empirical similarity function derived from the data, as a definition of objective probabilities. We hold that this is a reasonable definition in certain domains of application, such as described in Example 3. In Section 4, we compare our definition to alternative definitions that are based on statistical techniques. We argue that our approach is more appropriate, mostly because it is a natural extension of the frequentist approach, and because it is axiomatically based.

Yet, we do not view our definition as universally applicable. In fact, the axiomatizations of our formula are also helpful in identifying classes of situations in which it might be inappropriate. Example 4 is such a situation. We are not aware of any method for the assignment of objective probabilities that would be intuitive in situations such as Example 4. Our definition certainly isn't. We devote Section 5 to limitations of our approach. Finally, Section 6 discusses possible directions for extending our definition to a wider class of cases.

2 Similarity-Weighted Relative Frequencies

2.1 The formula

For concreteness, we stick to Example 3 in the exposition. Let the variable of interest be $Y \in \{0, 1\}$, indicating success of a medical procedure. The characteristics of patients are $X = (X^1, ..., X^m)$. These variables are real-valued, but some (or all) of them may be discrete. We are given a database

consisting of past observations of the variables $(X, Y) = (X^1, ..., X^m, Y)$, denoted $(X_i, Y_i)_{i \leq n}$. A new case is introduced, with characteristics $X_{n+1} = (X^1_{n+1}, ..., X^m_{n+1})$, and we are asked to assess the probability that $Y_{n+1} = 1$.

Assume that we are also equipped with a "similarity" function s such that, for two vectors of characteristics, $X_i = (X^1_i, ..., X^m_i)$ and $X_j = (X^1_j, ..., X^m_j)$, $s(X_i, X_j) > 0$ measures the similarity between a patient with characteristics X_i and another patient with characteristics X_j. The similarity function s will later be estimated from the data. We propose to define the probability that $Y_{n+1} = 1$, given the function s, by[3]

$$\hat{Y}^s_{n+1} = \frac{\sum_{i \leq n} s(X_i, X_{n+1}) Y_i}{\sum_{i \leq n} s(X_i, X_{n+1})}. \tag{1}$$

That is, the probability that Y_{n+1} be 1, i.e., that the procedure will succeed in the case of patient X_{n+1}, is taken to be the s-weight of all past successes, divided by the total s-weight of all past cases, successes and failures alike.

2.2 Intuition

Formula (1) is obviously a generalization of the notion of empirical frequency. Indeed, should the function s be constant, so that all observations are deemed equally relevant, (1) boils down to the relative frequency of $Y_i = 1$ in the database. If, however, one defines $s(X_i, X_j)$ to be the indicator function of $X_i = X_j$ (allowing for the value 0 in case the vectors differ from each other, and setting it to be 1 in case they are equal), then formula (1) becomes the conditional relative frequency of $Y_i = 1$, that is, its relative frequency in the sub-database defined by X_{n+1}. It follows that (1) suggests a continuous spectrum between the two extremes: as opposed to conditional relative frequencies, it allows us to use the entire database. This is particularly useful

[3]For simplicity, we assume that s is strictly positive, so that the denominator of (1) never vanishes. But in certain situations, such as conditional frequencies, one may wish to allow zero similarity values. Leshno (2007), who extends the axiomatization to this case, and allows a sequence of similarity functions that are used lexicographically as in (1).

in the medical example, where the database defined by $X_i = X_{n+1}$ may be very small or even empty. At the same time, it does not ignore the variables X, as does simple relative frequency over the entire database. Thus, formula (1) uses the entire database, but it still allows a differentiation among the cases depending on their relevance.

2.3 Axiomatic Derivations

Formula (1) has been axiomatized in Gilboa, Lieberman, and Schmeidler (GLS, 2006) for the case discussed here, namely, the estimation of the probability of a single event, or, equivalently, of the distribution of a random variable with two possible values. Billot, Gilboa, Samet, and Schmeidler (2005) provide an axiomatization of the same formula in the case that the random variable under discussion may assume at least three distinct values.[4] Gilboa, Lieberman, and Schmeidler (2007) extend this axiomatization to the assessment of a density function of a continuous variable. While these axiomatic derivations differ in the framework, as well as in the assumptions regarding which data are observable, they all use a "combination" axiom, which states, roughly, that if a certain conclusion should be arrived at given two disjoint databases, then this conclusion should also be the result of the union of these databases.[5]

The basic logic of the axiom is as follows. Assume that Mary asks her physician whether the operation is more likely to succeed than not. Suppose that the physician says that "chances are" it will, meaning that success is more likely than failure. Mary decides to seek a second opinion. She consults another doctor, who has been working in a different hospital for many years. Let us assume that both doctors have the same inference algorithm, and that they only differ in the databases they have been exposed to. Suppose that

[4] The axiomatization in Billot et al. (2005) relies on the fact that space of probability vectors has at least two dimensions, and it therefore cannot be adapted to the case or a binary variable (i.e., the one-dimensional case).

[5] Such an axiom was also used in Gilboa and Schmeidler (2001, 2003).

the second doctor also thinks that success is more likely than failure. Should Mary ask the two to get together and exchange databases?

If Mary does not feel that the two doctors should exchange data, she implicitly believes that a conclusion, which has been warranted given each of the two databases, will also be warranted given their union. Casual observation suggests that people are generally reassured when they find that the advice of different experts converge. Hence we find the basic logic of combination axiom rather natural.

Gilboa and Schmeidler (2003) show that several well-known statistical techniques satisfy the combination axiom. These include likelihood ranking by empirical frequencies, kernel estimation of a density function, kernel classification, and maximum likelihood ranking of distributions. The fact that all these techniques obey the same principle, namely the combination axiom, may be taken as an indirect piece of evidence that the axiom is a good starting point for a theory of belief formation. Having said that, there are several important classes of applications where the combination axiom is unreasonable. We discuss these in Section 5.

3 Empirical Similarity

3.1 Best fit

As mentioned above, the probability obtained from similarity-weighted frequencies depends on the similarity function one employs. Which function should we use? In an attempt to avoid arbitrary choices, and in the hope of retaining objectivity, we define the *empirical similarity* to be the similarity function that best explains the database, assuming that we use it as in (1). To simplify the estimation problem, we choose a particular functional form, and thus render the problem parametric. Specifically, we specify a vector of positive weights $w = (w_1, ..., w_m)$, consider the weighted Euclidean distance

corresponding to it,

$$d_w(\bar{x}, \bar{x}') = \sqrt{\sum_{j \leq m} w_j (x_j - x'_j)^2}$$

and use as a similarity function the negative exponential of the weighted distance:[6]

$$s_w = e^{-d_w}.$$

Given the database, for each vector w, one may calculate, for each $i \leq n$, the value

$$\hat{Y}_i^{s_w} = \frac{\sum_{j \neq i} s_w(X_i, X_{n+1}) Y_j}{\sum_{j \neq i} s_w(X_i, X_{n+1}))}. \tag{2}$$

The goodness of fit can be measured by

$$SSE(w) = \sum_{i \leq n} (\hat{Y}_i^{s_w} - Y_i)^2.$$

It then makes sense to ask, which vector w minimizes the sum of squared errors, $SSE(w)$. The minimizer of this function is then used in (1) to define the probability that $Y_{n+1} = 1$. When we use (1) in conjunction with the SSE-minimizing vector w, we obtain probability estimates that are "objective" in the same sense that classical statistics generally is: one may resort to statistical considerations for the choice of the general procedure, but no specific knowledge relating to the application is needed to implement the procedure. In Example 3, Mary needs to consult a statistician for the choice of the functional form of the similarity function, as well as for the measure of goodness of fit. But she does not need to consult a physician. Indeed, the resulting probability assessments are independent of the physician's subjective judgment or expertise. These assessments follow directly from the database, and they may serve an inexperienced doctor just as an experienced one.

[6] The exponential function was characterized in Billot, Gilboa, and Schmeidler (2005). They provide simple conditions on assessments, presumably generated by similarity-weighted averages, and show that these conditions are equivalent to the existence of a norm on $\mathbb{R}^m$ such that the similarity function between two vectors is the negative exponent of the norm of the difference between these vectors. The choice of the weighted Euclidean distances out of all possible norms is made for simplicity.

3.2 Comparison with the Notion of IID Random Variables

The textbook examples of classical statistics have to do with i.i.d. random variables, that is, random variables that are identically and independently distributed. These properties guarantee the laws of large numbers, the central limit theorem, and all the results that derive from these. Importantly, the laws of large numbers offer a natural definition of probability by relative frequency. Both the identical distribution and the statistical independence assumptions may be relaxed to a certain extent without undermining the laws of large numbers. But these assumptions cannot be dropped completely. If there is neither statistical independence, nor a certain weakening thereof, the size of the sample does not guarantee that the relative frequency would converge to anything at all, let alone to a number that can be interpreted as probability. Worse still, if the distributions of the random variable are not identical, or close to identical, it is not at all clear what is the "probability" that the relative frequency should converge to.

The mapping between the assumptions of i.i.d. observations in statistics and our model is not straightforward. Both notions, "identical" and "independent", are defined in probabilistic terms, whereas in our model no probability is assumed. Indeed, the model attempts to invest this concept with meaning, and therefore cannot assume it as primitive. Yet, our model suggests intuitive counterparts to these assumptions. Identicality of distribution is somewhat akin to identicality of the circumstances, i.e., of all observed variables x. One might argue that we cannot directly observe the distribution of the variable of interest, y, and if all observed variables x assume the same values, this is the closest that we can get to "identical distribution" in an empirical study. Stochastic independence of random variables is a rather strong condition, which implies the absence of causal relationships between the variables in question. This causal independence is implicit in our combination axiom.

Viewed from this perspective, our approach suggests that the independence assumption is, in a very vague sense, more fundamental than the identical distribution assumption. Our model drops the assumption that all observations are taken under identical conditions. In doing so, it foregoes the notion of a probability number that exists in some abstract or platonic sense, independent of our sample, and to which relative frequencies might converge. In our model, the probability of the event $y_t = 1$ occurring at observation t is a number that differs with t, depending of the x_t values. Our approach is therefore not an attempt to measure a quantity whose existence is external to the sample. Rather, our approach defines certain rules, by which the term "probability" can be used in an objective and well-defined way. And we argue that for this notion of "objective probability" one need not assume any notion of "identical repetition".

By contrast, our model heavily relies on the combination axiom, which may be viewed as retaining some notion of independence. Indeed, in the presence of causal dependencies neither our axioms nor our formula are very plausible. Thus, our approach suggests that in order to discuss objective probabilities, one need not resort to any notion of identical repetition, but one does need some notion of independence.

3.3 Statistical Theory

Finding the parameters that minimize the sum of squared errors is an accepted way of selecting the "best" model. But in order to employ statistical inference techniques such as hypotheses tests and confidence intervals, one needs to couch the similarity-weighted frequency formula in a statistical model. In GLS (2006) we analyze the following statistical model.

For $t = 2, ..., n$, we assume that

$$Y_t^{s_w} = \frac{\sum_{i<t} s_w(X_i, X_t) Y_i}{\sum_{i<t} s_w(X_i, X_t)} + \varepsilon_t \tag{3}$$

where $\varepsilon_t \sim N(0, \sigma^2)$, independently of the other variables.

In such a model it makes sense to ask whether the point estimates of the unknown parameters are significantly different from a pre-specified value, and in particular, from zero. In GLS (2006) we focus on maximum likelihood estimation of the parameters $(w_j)_j$, and we develop tests for such hypotheses. Observe, however, that the statistical model (3) differs from (2) in that, in the former, each observation is assumed to depend only on observations that precede it in the database. This assumes a certain order of the datapoints. When no such order is naturally given, such an order may be chosen at random. In this case, the statistical analysis should be refined to reflect this additional source of randomness.

4 Related Definitions

The problem of predicting a variable Y based on observable variables $X^1, ..., X^m$ is extensively studied in statistics, machine learning, and related fields. Among the numerous methods that have been suggested and used to solve such problems one may mention linear and non-linear regression, neural nets, linear and non-linear classifiers, k-nearest neighbor approaches (Fix and Hodges, 1951, 1952), kernel-based estimation (Akaike, 1954, Silverman, 1986, Scott, 1992), and others. Indeed, kernel-based methods are very similar to similarity-weighted frequencies.

When the variable of interest, Y, is binary (0 or 1), the approach that is probably the most popular in medical research is logistic regression. This method, introduced by McFadden (1974), uses the measurable variables in a linear formula, which is transformed in a monotonic way to a number between 0 and 1. This number can be computed for each given set of coefficients of the variables, and it is taken to be the predicted probability that the event in question will materialize in the next observation. Logistic regression finds the coefficients that result in the "best fit" that can be obtained between the predicted probabilities and the actual observations. This process may be

viewed as a possible definition of the term "probability" in Example 3.

The probability numbers generated by logistic regression depend on the predicting variables in these observations. At the same time, these probabilities can be thought of as "objective", because they do not resort to a physician's subjective assessments. Rather, they rely solely on observed data. Admittedly, statistics' claim to objectivity is always qualified. Different choices that a statistician makes in the estimation process will result in different outcomes. Yet, logistic regression, as well as the empirical similarity method we propose are objective in the sense that they do not require the statistician to consult with a medical expert in order to generate predictions or estimate probabilities.

It is generally expected that each method for the assessment of probabilities will be more successful in certain applications and less in others. Finding how well each method performs in a particular type of application is an empirical question that is beyond the scope of this paper. At the theoretical level, we hold that the method offered here has several advantages over the alternatives mentioned above. First, the similarity-weighted frequencies are an intuitive extension of simple frequencies, and can thus be offered as a definition of the notion "probability". By contrast, some of the alternative methods might be useful predicting tools, but they are not intuitively interpretable as "probability". Second, our basic formula (1) appears to be the only one which is axiomatically derived. Third, not all alternative method have the statistical theory required for statistical inference (in particular, hypotheses testing). Finally, in contrast to, say, logistic regression, our approach does not assume any functional relationship between the observed variables and the predicted variable. In this sense, the empirical similarity approach is more "epistemically humble", in that it allows the data to determine not only the actual probability assessment, but also the way that this assessment is computed (via determination of the similarity function).

5 Limitations

Gilboa and Schmeidler (2003) contain an extensive discussion of the combination axiom, including an attempt to characterize several classes of counter-examples, that is, of situations in which the axiom appears unreasonable. We will not replicate this discussion here, but we will mention a few classes of problems in which one should not expect the axiom to hold, and, consequently, one should not use the methods that are restricted to satisfy it.

The first class of counter-examples involves theorizing, that is, inductive inferences from cases to underlying theories, and then deductive inference back from these theories to future cases. For example, when one uses past observations to learn the parameter of a coin, p, and then uses probability theory to make predictions regarding sequences of tosses of that coin using the best estimate of p, the combination axiom is unlikely to hold.

Similarly, if one believes that the observations are generated by a linear function, uses the data to estimate a linear regression formula, and then uses the estimated formula to make predictions, one is unlikely to satisfy the combination axiom. Indeed, it is evident from the basic similarity-weighted average formula (1) that it does not make any attempt to identify trends. To consider an extreme example, assume that $m = 1$ and that the database contains many points with $x = 1$, for which the relative frequency of $Y = 1$ was .1. There are also many points with $x = 2$ and a relative frequency of .2 for $Y = 1$, and so on for $x = 3$ and $x = 4$. Next assume that we are asked to make a prediction for a new case in which $x_t = 5$. It seems patently plausible to suggest that the probability that $Y_t = 1$ be 0.5. It makes sense to identify a trend, by which the probability of $Y = 1$ goes up with x. In fact, logistic regression seeks precisely this kind of relationships. But the similarity-weighted average method will fail to produce a value that is outside the observed range of $[0.1, 0.4]$. It is important to recall that the similarity-weighted average does not seek trends, and, more generally, does not attempt to theorize about the data. It engages only in case-to-case induction, but not

in case-to-rule induction coupled with rule-to-case deduction.

The second class of counter-examples to the combination axiom involves changes in the similarity function. In particular, if the probabilistic reasoner perform so-called second-order induction (Gilboa and Schmeidler, 2001), that is, if she learns which similarity function should be used to learn from past cases about the future, then the combination axiom is again an unlikely principle. Learning of the similarity function may involve qualitative insights, if, say, a physician, after examining a large database, says, "and it suddenly dawned on me that the common feature to all these cases was...". But such learning may also be quantitative and follow from purely statistical reasons: with the accumulation of more data, the need to use more remote observations is reduced, and one may obtain better assessments by restricting attention to close cases. More generally, learning of the similarity function, that is, any process that adapts the similarity function as a function of the data should be expected to violate the combination axiom.

Obviously, our approach faces a difficulty here. On the one hand, we justify the similarity-weighted average based on the combination axiom, which is violated when the similarity function is learnt itself. On the other hand, the notion of empirical similarity is precisely one of learning the similarity function. Thus, a probabilistic reasoner who would follow our advice to compute the empirical similarity will thereby violate our recommendation to satisfy the combination axiom, and will consequently not be sure that the similarity-weighted average method makes sense to begin with.

One possible resolution is to assume that the similarity function is updated only at certain periods, and between each two such consecutive periods the combination axiom holds. We conjecture that the axiomatizations mentioned above would have approximate counterparts with bounded databases, which would allow the use of formula (1) between updating periods. Admittedly, this resolution is rather awkward, and more elegant axiomatizations of similarity-weighted frequencies with a similarity function that is learned are

called for.

6 Future Directions

The discussion above suggests several directions in which one may extend the empirical similarity approach to the definition of probabilities. First, one should have more satisfactory theories, allowing the similarity function to be learnt and refined in the process, in a way that parallels the choice of a kernel function in non-parametric estimation. (See Silverman, 1986.) Second, analogical, case-based reasoning which is incorporated in similarity-weighted frequencies should be combined with deductive, rule-based reasoning. For instance, one may extend the similarity-weighted formula so that each observation (x_i, y_i) will give support not only to the value observed y_i, but also to various functions $f(x)$ that the observation approximately satisfies, i.e., to functions f such that $y_i \approx f(x_i)$. Thus, if all points observed lie near the graph of a certain function f, this function will gain support from each of the observations, and will thus offer itself as a natural generalization of the cases to a rule. The class of functions f one allows into this analysis has to be limited to make the analysis meaningful (and to avoid "overfitting" by finding a function that matches all the data precisely, but that does poorly in prediction). It is a challenge to find natural limitations on the class of functions that will retain a claim to objectivity.

Another extension might combine Bayesian reasoning with similarity-weighted averages. One may start with a Bayesian network (see Pearl, 1986), reflecting possible dependencies among variables, and assess probabilities on each edge in the network by the empirical similarity technique. These probability numbers will then be used by the Bayesian network to generate probabilistic predictions that make full use of the power of Bayesian reasoning. In this case, again, part of the challenge is to find ways to develop Bayesian networks that will be "objective".

However, it is not at all obvious that these and other extensions, or, in fact, any other approach, can come up with a reasonable definition of objective probability in Example 4 above. The main difficulty appears to be the causal dependence between cases. When cases are causally independent, an observation of one case may teach us something about the likelihood of the occurrence of an event in another case. As long as the latter is a fixed target, one may have a hope that, with sufficiently many observations, one may learn more about this likelihood, to a degree that it can be quantified in a way that most people would agree on. But when causal dependence is present, an observation of a particular case not only reveals information about another case, it also changes its likelihood. Thus, we are after a moving target, and find it difficult to separate the process of observation from the process observed.

Idiosyncrasy of cases and causal relationships do not allow us to define objective probabilities by empirical frequencies. These two phenomena also make it difficult to assign observable meaning to counterfactual propositions. It is relatively easy to understand what it meant by the statement "If I were to drop this glass, it would break". There are many cases of practically identical glasses being held and being dropped, and since these cases are assumed to be causally independent, this counterfactual statement has a verifiable meaning. Correspondingly, one can design an experiment that will be viewed as a test of this statement. By contrast, it is much harder to judge the veracity of the counterfactual, "Had Hitler crossed the channel, he would have won the war". First, historical cases of war are never identical. Second, they are seldom causally independent. Our approach suggests that a theory of counterfactuals might more easily deal with the first problem than with the second.

At present, we are not convinced that the notion of objective probability can be meaningfully defined in situations involving intricate causal relationships between observations. However, it seems obvious that when causal

independence holds, the definitions discussed here can be greatly improved upon.

7 References

Akaike, H. (1954), "An Approximation to the Density Function", *Annals of the Institute of Statistical Mathematics*, **6**: 127-132.

Bernoulli, J. (1713), Ars Conjectandi. 1713.

Billot, A., I. Gilboa, D. Samet, and D. Schmeidler (2005), "Probabilities as Similarity-Weighted Frequencies", *Econometrica*, 73, 1125-1136.

Billot, A., I. Gilboa, and D. Schmeidler (2005), "Exponential Similarity", mimeo.

de Finetti, B. (1937), "La Prevision: Ses Lois Logiques, Ses Sources Subjectives", Annales de l'Institute Henri Poincare, 7: 1-68.

Ellsberg, D. (1961), "Risk, Ambiguity and the Savage Axioms", Quarterly Journal of Economics, 75: 643-669.

Fix, E. and J. Hodges (1951), "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties". Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

——— (1952), "Discriminatory Analysis: Small Sample Performance". Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

Gilboa, I., O. Lieberman, and D. Schmeidler (2006), "Empirical Similarity", *Review of Economics and Statistics*, forthcoming.

——— (2007), "Similarity Based Prediction", mimeo.

Gilboa, I., A. Postlewaite, and D. Schmeidler (2006), "Rationality of Belief", mimeo.

Gilboa, I. and D. Schmeidler (1995), "Case-Based Decision Theory", *Quarterly Journal of Economics*, **110**: 605-639.

——— (2001), *A Theory of Case-Based Decisions*, Cambridge: Cambridge University Press.

——— (2003), "Inductive Inference: An Axiomatic Approach", *Econometrica*, 71, 1-26.

Hume, D. (1748), *Enquiry into the Human Understanding*. Oxford, Clarendon Press.

Leshno, J. (2007), Similarity-Weighted Frequencies with Zero Values, mimeo.

Machina, M. and D. Schmeidler (1992), "A More Robust Definition of Subjective Probability", *Econometrica*, 60: 745-780.

McFadden, D. (1974), "Conditional Logit Analysis of Qualitative Choice Behavior". In *Frontiers in Econometrics*, ed. P. Zarembka, 105–142. New York: Academic Press.

Pearl, J. (1986), "Fusion, Propagation, and Structuring in Belief Networks", *Artificial Intelligence*, 29: 241–288.

Pearl, J. (2000), *Causality*. Cambridge University Press.

Ramsey, F. P. (1931), "Truth and Probability", in The Foundation of Mathematics and Other Logical Essays. New York, Harcourt, Brace and Co.

Rostek, M. (2006), "Quantile Maximization in Decision Theory", mimeo.

Savage, L. J. (1954), *The Foundations of Statistics*. New York: John Wiley and Sons. (Second addition in 1972, Dover)

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley and Sons.

Shafer, G. (1986), "Savage Revisited", *Statistical Science*, 1: 463-486.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*. London and New York: Chapman and Hall.